SCIENTIFIC REPORTS
natureresearch

OPEN

Learning leads to bounded rationality and the evolution of cognitive bias in public goods games

Olof Leimar[1*] & John M. McNamara[2]

In social interactions, including cooperation and conflict, individuals can adjust their behaviour over the shorter term through learning within a generation, and natural selection can change behaviour over the longer term of many generations. Here we investigate the evolution of cognitive bias by individuals investing into a project that delivers joint benefits. For members of a group that learn how much to invest using the costs and benefits they experience in repeated interactions, we show that overestimation of the cost of investing can evolve. The bias causes individuals to invest less into the project. Our explanation is that learning responds to immediate rather than longer-term rewards. There are thus cognitive limitations in learning, which can be seen as bounded rationality. Over a time horizon of several rounds of interaction, individuals respond to each other's investments, for instance by partially compensating for another's shortfall. However, learning individuals fail to strategically take into account that social partners respond in this way. Learning instead converges to a one-shot Nash equilibrium of a game with perceived rewards as payoffs. Evolution of bias can then compensate for the cognitive limitations of learning.

Many different cognitive processes fall under the heading of learning. The most basic is when an individual learns solely from rewards, without forming a more sophisticated cognitive model of the situation. This corresponds to the much-studied learning processes in classical and operant conditioning in animal psychology[1], as well as to the standard, model-free approach to reinforcement learning in the study of machine learning[2]. It is this kind of learning we investigate here, where individuals explore through randomness in their actions and come to prefer actions that result in higher than so-far estimated rewards.

In social interactions, individuals typically vary in their characteristics in ways that influence costs and benefits. Examples could be differences in size and strength in aggressive interactions and variation in individual quality in cooperative interactions[3,4]. Variation in quality can cause individuals to vary in their investments into a joint project, which in turn can have the consequence that social partners respond through changes in their own investments. A question we raise is whether reinforcement learning allows individuals to take such dynamic responses from social partners into account when adjusting their own investments. As we show, the answer to the question can be no, because the responses by social partners occur over a too long time scale to be captured by learning. Instead, we show that the investment outcome of reinforcement learning in repeated rounds of the game corresponds to a Nash equilibrium of a one-shot game with the rewards acting as payoffs that are known to all players. Such a property of learning being myopic to future consequences of current actions can be seen as a kind of bounded rationality[5,6]. The phenomenon leaves open the possibility that evolutionary changes in the perceived rewards instead adjusts behaviour in a way that takes into account responses by social partners. The process can be thought of as an evolution of a bias in the innate perception of rewards, referred to as primary rewards or reinforcements in animal psychology. We show that such an evolution of cognitive bias indeed can occur, through the evolution of a tendency for individuals to act as if they underestimate their own quality, entailing an overestimation of their Darwinian fitness cost of investing into a project. The net effect is a lowering of investments compared to what would be the case for a Nash equilibrium of a one-shot game where individuals know the qualities of all players.

[1]Department of Zoology, Stockholm University, SE-106 91, Stockholm, Sweden. [2]School of Mathematics, University of Bristol, Bristol, BS8 1UG, UK. *email: olof.leimar@zoologi.su.se

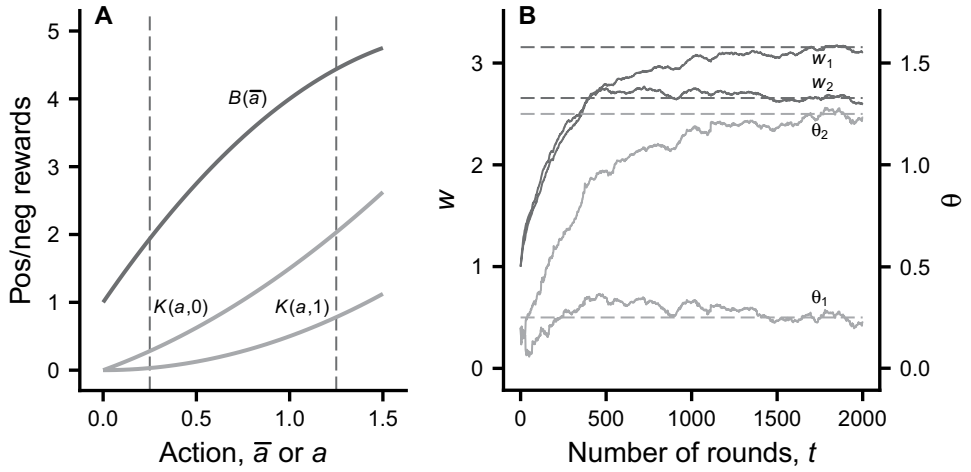

Figure 1. Illustration of the learning model. Panel (A) illustrates the benefit and cost as functions of the investment actions. The two curves for the cost correspond to qualities $q=0$ and $q=1$. See Eqs (2, 3) for the formulas. Panel (B) shows simulated learning dynamics of the estimated values w_i and mean actions θ_i for an interaction between two individuals with qualities $q_1=0$ and $q_2=1$. The dynamics of w_i and θ_i are given in Eqs (7, 10). The starting point of learning was (arbitrarily) chosen as $w_i=1.0$ and $\theta_i=0.2$. The dashed lines are one-shot game predictions for the estimated value w_i and the mean investment θ_i, corresponding to the investments in Eq. (14). These values of θ_i are also indicated in panel (A). Parameter values are: $g=2$, $B_0=1$, $B_1=4$, $B_2=-2$, $K_1=1$, $K_{11}=1$, $K_{12}=-1$, $\sigma=0.05$, $\alpha_w=0.04$, and $\alpha_\theta=0.002$.

Our analysis is inspired by McNamara *et al.*[7], who studied negotiation rules in games of cooperation with continuous actions. Our approach is to let reinforcement learning give rise to a "negotiation rule", and then to examine the evolutionary consequences of such a rule. We study a public goods game where in each round each group member invests an amount into a joint project and shares equally in the benefit of the total investment by the group. Over the rounds, individuals learn to adjust their investments. For the learning dynamics, we use the actor-critic approach to reinforcement learning[2], which is similar to so-called Bush-Mosteller learning[8]. We use a combination of analytical derivation and individual-based simulation to reach our main conclusion, that cognitive bias evolves as a consequence of the bounded rationality of learning. In summary, we show that if learning is driven by short-term rewards, cognitive biases may evolve as a compensating mechanism.

Results

Model overview. In each generation there are a number of investment rounds, $t=1, \ldots, T$, with an investment game involving a group of individuals. A group of size g stays together for life and a_{it} is the investment by individual i in round t. Each game is independent and has the same payoff structure, and group members can learn about the rewards (payoffs) from the successive rounds. Group members can differ in individual quality q_i, which influences the cost of investment. The quality is a non-genetic aspect of an individual's phenotype that influences its capacity to invest. The qualities are assumed not to vary between rounds of the game, but an individual's quality is drawn randomly from a distribution at the start of a generation.

Concerning what is "known" by group members, we assume that they do not have any particular information, including about their own quality, but that they learn about which investment to make through the rewards they receive. We thus assume that at the start of a generation individuals do not have information about any of the q_i in the group, and during the interaction they perceive their own rewards. This situation corresponds to traditional instrumental or operant conditioning, but in a game situation. The net reward for individual i from round t is a benefit B, which depends on the group mean investment, minus a cost K, which depends on the individual's own investment and quality (Fig. 1A and Eqs 1–4). We first assume that payoffs are perceived as rewards by the players. To study the evolution of cognitive bias, we than investigate whether individuals could evolve to perceive rewards that differ from the payoffs that correspond to Darwinian fitness.

Actor-critic learning. We implement the repeated investment game as a reinforcement learning process, using the actor-critic method described in sections 13.5–13.7 of[2], for the case without state transitions (only one state). Individuals learn which actions to use from the rewards they perceive. They use a temporal difference (TD) method to update a value w_{it} (estimated value by individual i at the start of round t), involving a TD error, or prediction error, which is the difference between actual and estimated rewards. The prediction error can be thought of as a reinforcement. Individuals select actions using a policy, expressed as a probability density $\pi(a|\theta_{it})$ of using the action a, assumed to be normal with mean θ_{it} and standard deviation (SD) σ. A so-called policy gradient method (ch. 13 in[2]) is used to update the parameter θ_{it}, representing the mean investment action. In the learning process, the w_{it} and θ_{it}, $i=1, \ldots, g$, then perform a random walk in a $2g$-dimensional space (Fig. 1B), specified by Eqs 5–11; (see Methods).

Reinforcement learning based on a policy gradient is thought to have good convergence properties (e.g., ch. 13 in[2]), in the sense that for small rates of learning a local optimum is approached. In a game situation, the outcome

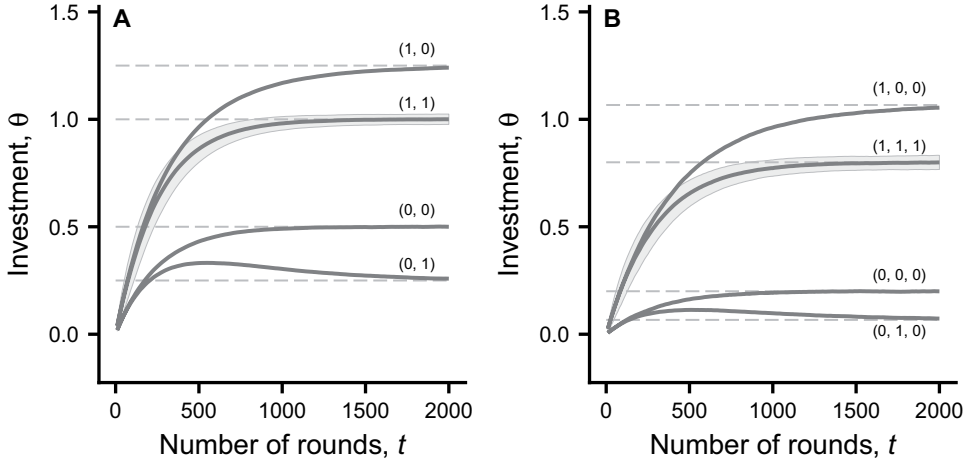

Figure 2. Mean and SD of simulated investment actions for individual $i = 1$ in populations of groups, plotted over the rounds of learning. At the start of learning, individuals are assigned random qualities from the set {0, 1} and the curves are labelled with with the qualities, q_i, $i = 1, \ldots, g$, of individuals in a group. The spread (SD) of values of θ in the population is shown as grey shading only for the subset of groups where all $q_i = 1$ (for clarity, to avoid overlap). Panel (A) shows all cases of group compositions with $g = 2$, namely groups with $q_1 = 1, q_2 = 1$; $q_1 = 1, q_2 = 0$; $q_1 = 0, q_2 = 1$; and $q_1 = 0, q_2 = 0$. Panel (B) shows a subset of cases of group compositions with $g = 3$, labelled q_1, q_2, q_3. The total population size is 24 000 individuals in both panels. The dashed lines are one-shot game predictions, from Eq. (14). Other parameters are as in Fig. 1.

of learning in successive rounds might approximate a Nash equilibrium of a one-shot game with the rewards as payoffs. In this one-shot game the payoffs, including the dependence on individual qualities, are known to the players, and are given by Eq. (4). From our individual-based simulations (Figs 1B and 2), the learning dynamics approach this Nash equilibrium, which is specified by Eqs (13, 14). Because learning is a stochastic process, driven by the individual exploratory choices of investment actions, there is variation in learning trajectories between groups with identical compositions of qualities. This variation is shown as shading, indicating ± 1 SD, in Fig. 2. For small rates of learning we also show that the learning dynamics is approximately a vector autoregressive process[9] around the Nash equilibrium (see SI, Figs S1 and S2).

Evolution of cognitive bias. We found that the learning outcome corresponds to a Nash equilibrium of a one-shot game, with payoffs illustrated in Fig. 1A and specified in Eq. (4). For these payoffs, the cost of an action depends on the "true" quality q_i of a player. However, the analysis of learning applies in the same way if the qualities q_i are replaced by "perceived qualities" p_i, as in Eq. (15), meaning that individual i behaves as if its quality is p_i. We refer to the rewards used in learning as "perceived rewards". An individual of quality q_i would then learn from rewards corresponding to its perceived quality p_i, which might differ from q_i. Note that we assume that individuals only perceive their benefit and cost in each round. The bias thus occurs in an individual's perception of its cost of investment, but for convenience we express it as a bias in perceived quality. Specifically, an individual is assumed to perceive a cost that corresponds to its perceived quality p_i, while its Darwinian fitness cost is given by its true quality q_i. We define an individual's cognitive bias as the difference between its perceived and true qualities: $d_i = p_i - q_i$. We also assume that perceived qualities satisfy $p_i \leq 1$ and can be negative. This allows d_i to be either positive or negative, with negative d_i corresponding to higher perceived costs of investment.

A main result of our analysis is that when there are social partners, i.e. for group size $g > 1$, zero cognitive bias, i.e. $d_i = p_i - q_i = 0$, is not an evolutionary equilibrium, but instead a negative bias evolves. An intuitive explanation is that, given the perceived qualities of the group members, learning approaches a one-shot Nash equilibrium for these perceived qualities. The learning outcome does not strategically take into account that social partners respond to an individual's lowered investment by increasing their investments somewhat. From the definition of a Nash equilibrium, it then follows that the individual can gain fitness by having a cognitive bias, i.e., by lowering its perceived quality from $p_i = q_i$. In effect, an individual whose perceived quality is lower than the real quality makes smaller investments, which in turn means that other players end up making larger investments. The individual thus makes a fitness gain from the biased perception. The derivation of this result appears in the Methods, Eq. (16), and the evolutionarily equilibrium bias is given in Eq. (17), with detailed derivation in SI. This result is illustrated in Fig. 3A, which shows the evolution of a genetically determined perceived quality p_i in a population where all individuals have true quality $q_i = 1$. As can be seen, a negative cognitive bias $d_i = p_i - q_i$ evolves.

For a given composition of q_i, $i = 1, \ldots, g$, in a group one can find the evolutionarily stable perceived qualities p_i^* using Eq. (16). For the benefit and cost functions in Eqs (2, 3), which we use for illustration, this simplifies to Eq. (17), from which it follows that the evolutionarily stable cognitive bias $d_i^* = p_i^* - q_i$ depends on the group average quality $\bar{q}$. However, it is not reasonable to assume that an individual has an evolved innate underestimation of its true quality that depends on the particular group composition, because this composition is not known to the individual at the start of a generation. Instead, in individual-based simulations we assume that the trait that

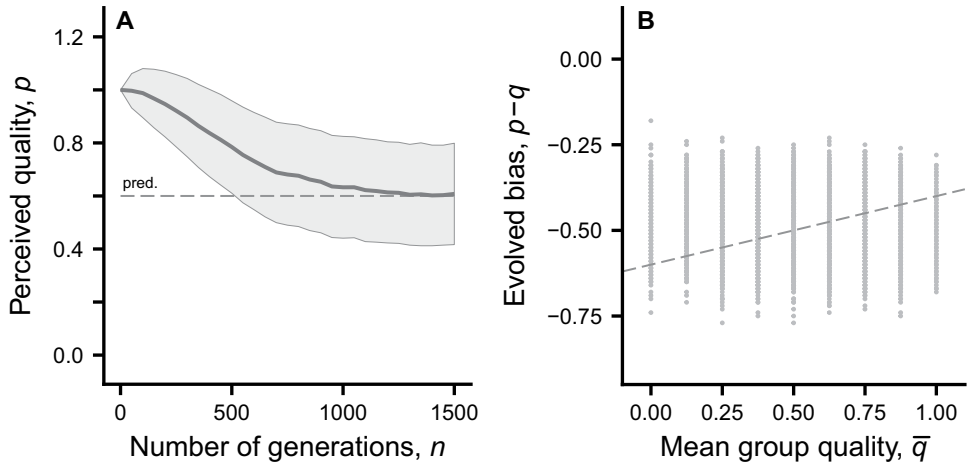

Figure 3. Illustration of the evolution of individual perceived quality p_i, through the genetically determined cognitive bias $d_i = p_i - q_i$, from individual-based simulation of populations similar to those illustrated in Fig. 2. Panel (A) shows evolution of mean and SD of $p_i = q_i + d_i$ over the generations in a population with groups of size $g = 2$, with true qualities $q_1 = 1$, $q_2 = 1$ (i.e., all individuals have true quality 1). The mutation rate for alleles for d_i is 0.05 and the mutant increment is normally distributed with an SD of 0.04. The dashed line is the prediction from Eq. (17). Panel (B) shows the bias d_i, as a function of the mean quality $\bar{q}$ in the group, for a population with groups of size $g = 2$ and with true qualities selected randomly at the start of a generation from the set {0.00, 0.25, 0.50, 0.75, 1.00}. The dashed line shows the prediction from Eq. (17) for each group composition in the final simulated generation. The mutation rate per allele for d_i is 0.001 with SD of mutant increments of 0.04. Other parameters are as in Fig. 1.

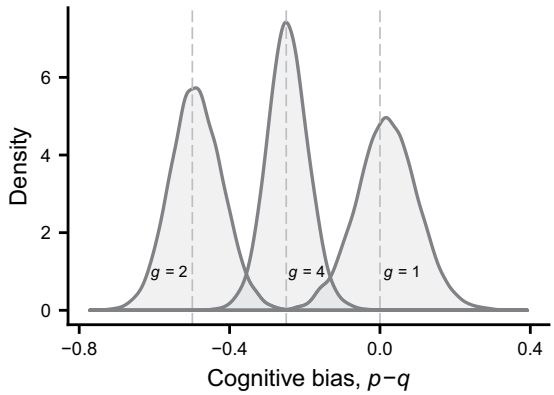

Figure 4. Illustration of the distribution of evolved cognitive bias $d_i = p_i - q_i$ for different cases of group sizes. Parameters are as in Fig. 3B, and the distribution for $g = 2$ comes from the population illustrated in Fig. 3B. The dashed lines give the prediction from Eq. (17), averaged over the different group compositions in the population.

evolves is simply a bias d_i, such that the perceived quality is $p_i = q_i + d_i$, irrespective of the kind of group the individual is a member of. An example with $g = 2$ and variation in true quality in the population appears in Fig. 3B. Our assumption means that perceived qualities cannot match the prediction from Eq. (17) for each particular group composition (Fig. 3B), but there is agreement between the population averages of the evolved and predicted cognitive biases (equal to −0.49 and −0.50, respectively).

This is further illustrated in Fig. 4, showing the outcome of individual-based simulations for populations with different group sizes. The most extreme bias occurs for $g = 2$, and as the group size g becomes large, the bias approaches zero (see SI). For solitary investing individuals ($g = 1$), there is no bias on average.

Discussion

A major conclusion from our analysis is that when individuals in a group learn how much to invest in a public goods game, there is scope for the evolution of cognitive bias, corresponding to an evolution of the perceived cost of investment into the public good (Figs 3 and 4). The reason is that cognitive limitations of reinforcement learning prevent individuals from fully taking into account how social partners respond to variation in the ability of individuals to invest. Reinforcement learning is a mechanism driven by immediate rewards, without foresight about the medium-term outcome of learning. This aspect of learning can be seen in Figs 1B and 2, where the

mean investment θ of the lowest-quality individual in a group approaches its equilibrium by first overshooting the eventual equilibrium value. Furthermore, the learning interaction is particularly beneficial for a low-quality individual (Fig. 1B), who ends up investing little, when interacting with higher-quality partners who end up investing more, by learning to compensate for the shortfall. This explains why evolutionary changes are in the direction of a reduced perceived quality, i.e. a negative cognitive bias.

For an individual to learn about how social partners respond to variation in its tendency to invest, several interactions with different social groups would be needed, where the individual could explore the consequences of changes in its tendency to invest. Even so, for an individual to learn that lowering its current investment increases rewards in future rounds, because others learn to increase their investments, the individual must connect current behaviour to future rewards. Animal psychology has shown that this can be difficult to do, in particular without any indicators to the individual that there might be such a causal connection. Pavlov[10] discovered that the time interval between conditioned and unconditioned stimuli (the CS-US interval) needs to be short for an association to be formed. Exceptions to this rule represent special adaptations, of which taste aversion learning is the best known[1]. It has also been shown that learning can occur for longer CS-US intervals, if the CS is highly salient and there are no interfering stimuli during the interval[1]. A clearly perceived chain of states and actions, leading to a goal, can also support more sophisticated learning about future consequences of current actions[11], but learning about social partners does not have a structure of that kind. It thus seems reasonable that unless individuals have some other special preparedness to connect current behaviour to medium-term rewards, mediated through the responses of social parters, this will be difficult to learn.

Game dynamics and learning. As described by Weibull in the proceeding from a Nobel seminar[12], the general idea that players of a game are members of populations and revise their strategies in a more-or-less myopic fashion was introduced in unpublished work by John Nash. This is now a foundation for game theory in economics[13–15], and has also been used for game theory in biology[16–18]. Game dynamics based on reinforcement learning, including the actor-critic method[2], can be seen as a variant of this approach, with its learning mechanisms inspired by experimental psychology and neuroscience. Thus, the TD updating of an estimated value[2], described in Eqs (6, 7), represent the critic component of an actor-critic mechanism and is connected to the influential Rescorla-Wagner model of classical conditioning[19] as well as to the reward prediction error hypothesis of dopaminergic neuron activity[20]. For the actor component, from Eqs (10, 11), changes in the tendencies to perform actions depend on the covariance of eligibility and reward. This learning mechanism has been given an interpretation in terms of synaptic neural plasticity[2,21]. It is worth noting that there is a certain similarity between the actor-critic learning dynamics in Eq. (11) and the so-called Price equation for selection dynamics[22]. Although these equations describe fundamentally different processes, natural selection vs. actor-critic learning, they are both helpful in providing intuitive understanding.

Bounded rationality. The cognitive limitations of learning have been put forward as an important reason for bounded rationality[6,23,24] and our work gives further support to the idea. It is a general principle that certain aspects of the situation an individual finds itself in might be learnt very slowly or not at all, even though they could influence payoffs. In our model, the effects on rewards of responses of social partners, resulting from learning about an individual's characteristics, do not influence the learning of investment actions. Instead, we found a learning outcome where investments converged on a Nash equilibrium of a one-shot game with perceived rewards as payoffs, even though group members stayed together over successive investment rounds and, in principle, might have discovered how social partners learn about investment variation.

Evolution of cognitive bias. The possibility of cognitive bias in decision making has been of interest in economics, psychology and biology. Among the examples are the base rate bias[25] and the judgement bias[26]. The general question of how to formulate an evolutionary theory of cognitive bias has also been raised[27].

An insight from our analysis is that the bounded rationality of learning leaves scope for evolution to adjust the rewards (primary rewards or preferences) in a way that corresponds to a cognitive bias in an individual's perception of its quality. With such a bias, learning by individuals results in an approach towards evolutionarily optimal behaviour. Our result is related to the idea of an "indirect evolutionary approach" in economic game theory[28,29], where players are assumed to know or learn about each other's preferences and to play a Nash equilibrium given the preferences, which are then assumed to be shaped by evolution. The connection with our work is that we showed that learning causes the investments to approach a one-shot Nash equilibrium given the perceived qualities, and the indirect evolutionary approach assumes that players know or find out each other's preferences and play a Nash equilibrium given these preferences.

A widespread and successful idea in animal psychology is that evolution causes primary rewards to indicate Darwinian fitness[1]. More generally, it is a basic element of evolutionary biology and behavioural ecology that actions can be given a Darwinian currency, in the form of reproductive value[30,31]. Our work here, as well as related work in economic game theory[29,32,33], shows that an exact correspondence between primary rewards and reproductive value need not hold. In our model this happened because of cognitive limitations of learning, although reproductive value was still important for the analysis.

As illustrated in Figs 3 and 4, there is variation between individuals in their cognitive bias, i.e. in how much their perceived qualities deviate from the true qualities, which is a consequence of a balance between selection, mutation and genetic drift. This is reminiscent of animal personality variation[34], where individuals differ in important behavioural characteristics. One often assumes that disruptive selection lies behind personality variation[35], but our results here show that there can be substantial variation also with stabilising selection on the trait in question. In general, whether selection is stabilising or disruptive, we propose that bounded rationality, from cognitive limitations of learning, opens up a possibility for individuals to vary in their characteristics, including cognitive biases in social interactions.

Methods

Model details. In round t, the group mean investment $\bar{a}_t$ is

$$\bar{a}_t = \frac{1}{g}\sum_{i=1}^{g} a_{it}. \tag{1}$$

The benefit of investment for each group member is assumed to be a concave, smooth function of the group average investment $\bar{a}$, having a negative second derivative. For illustration we use the special case

$$B(\bar{a}) = B_0 + B_1\bar{a} + \frac{1}{2}B_2\bar{a}^2, \tag{2}$$

where $B_1 > 0$ and $B_2 < 0$ (Fig. 1A). Maximum benefit occurs for $\bar{a} = -B_1/B_2$, and we might constrain actions to be smaller than this, to ensure that benefits increase with the actions. The cost $K(a_i, q_i)$ of investment a_i by group member i is assumed to be a smooth convex and increasing function of a_i that increases more rapidly with a_i for smaller q_i, and has a positive second derivative with respect to a_i. For illustration we use

$$K(a_i, q_i) = K_1 a_i + \frac{1}{2}K_{11}a_i^2 + K_{12}a_i q_i, \tag{3}$$

with $K_1 > 0$, $K_{11} > 0$ and $K_{12} < 0$ (Fig. 1A). We thus have a public goods game in each round with the payoff to player i given by

$$W_i(a_i, a_{-i}, q_i) = B(\bar{a}) - K(a_i, q_i), \tag{4}$$

where a_{-i} denotes the vector of actions of all individuals in the group except for i.

Reinforcement learning: the actor-critic approach. Actions are independent and normally distributed with mean θ_{it} and SD σ:

$$\pi(a|\theta_{it}) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(a-\theta_{it})^2}{2\sigma^2}\right). \tag{5}$$

For simplicity, we keep σ constant and rather small, but we note that variation in a is needed for a learner to explore and thus to discover how actions can be improved. Keeping with reinforcement learning notational conventions, the reward from Eq. (4) for individual i from the play in round t is denoted $R_{i,t+1}$. The TD error is given by

$$\delta_{it} = R_{i,t+1} - w_{it} = W_i(a_{it}, a_{-it}, q_i) - w_{it}. \tag{6}$$

This is used to update the learning parameter w_{it} as follows:

$$w_{i,t+1} = w_{it} + \alpha_w\delta_{it}, \tag{7}$$

where α_w is a learning rate parameter (we do not use discounting in our formulation of learning and each round is treated as a new episode[2]). The expected change in w_{it} is

$$\mathrm{E}\,[w_{i,t+1} - w_{it}|w_{\cdot t}, \theta_{\cdot t}] = \alpha_w \mathrm{E}\,[\delta_{it}|w_{\cdot t}, \theta_{\cdot t}]. \tag{8}$$

For the actor-critic method the learning updates for the policy involve the derivative of the logarithm of $\pi(a|\theta)$ with respect to θ, given by

$$\zeta_{it} = \frac{\partial \log\pi(a|\theta_{it})}{\partial\theta_{it}} = \frac{a-\theta_{it}}{\sigma^2}, \tag{9}$$

which sometimes is referred to as an eligibility. The update to the learning parameter θ_{it} is

$$\theta_{i,t+1} = \theta_{it} + \alpha_\theta\delta_{it}\frac{\partial \log\pi(a|\theta_{it})}{\partial\theta_{it}} = \theta_{it} + \alpha_\theta\delta_{it}\zeta_{it}, \tag{10}$$

where α_θ is a learning rate parameter. It is worth noting that the expectation of the increment in θ_i is proportional to the covariance of the TD error and the eligibility:

$$\mathrm{E}\,[\theta_{i,t+1} - \theta_{it}|w_{\cdot t}, \theta_{\cdot t}] = \alpha_\theta\,\mathrm{Cov}\,[\delta_{it}, \zeta_{it}|w_{\cdot t}, \theta_{\cdot t}]. \tag{11}$$

A frequent issue for actor-critic reinforcement learning is how the learning rates α_w and α_θ should be chosen. Learning involves changes in both the estimated value w_i and the action mean value θ_i and both are driven by the TD error δ_i. From Eqs (7, 10), noting that $a_i - \theta_i$ has a magnitude of about σ, we ought then to have

$$\alpha_w\Delta\theta \sim \frac{\alpha_\theta}{\sigma}\Delta w \tag{12}$$

for learning to cause the w_i and θ_i to move over approximate ranges Δw and $\Delta\theta$. We have used this relation in our learning simulations, with ranges Δw and $\Delta\theta$ of around 1.

Intuitively, from Eqs (6–10) we might expect $\partial W_i/\partial a_i = 0$ to hold approximately for a learning equilibrium. This would correspond to a Nash equilibrium, and is the motivation for the following analysis.

One-shot game. By our assumptions about the payoffs, this is a concave game, and using a result in[36] one can show that the game has a unique Nash equilibrium (see SI). This equilibrium should satisfy $\partial W_i/\partial a_i = 0$ or, from Eq. (4),

$$\frac{1}{g}B'(\bar{a}^*) = \frac{\partial K(a_i^*, q_i)}{\partial a_i}, \tag{13}$$

for $i = 1, \ldots, g$. It follows that $\partial K(a_i^*, q_i)/\partial a_i = \partial K(a_j^*, q_j)/\partial a_j$ and, because $\partial K(a_i^*, q_i)/\partial a_i$ is increasing in a_i and decreasing in q_i, that $a_i^* > a_j^*$ when $q_i > q_j$, so higher-quality individuals invest more at the equilibrium. Furthermore, using results in[37], one can show that a_i^* increases with q_i and decreases with q_j, $j \neq i$ (see SI, Equation S11). For our special case of Eqs (2, 3), one readily finds that

$$a_i^* = e_0 + e_1 q_i + e_2 \sum_{j \neq i} q_j = e_0 + e_1 q_i + e_2 (g-1)\bar{q}_{-i}, \tag{14}$$

where, for $g > 1$, $\bar{q}_{-i}$ is the average quality of all individuals the group except for i (see SI, Equation S21, for the coefficients). For large g we see from Eq. (13) that the equilibrium is for individual i to minimize $K(a_i, q_i)$.

Evolution of cognitive bias. The cost $K(a_i, q_i)$, from Eq. (3), is assumed to be the true cost of investment, measured in terms of Darwinian reproductive value, for an individual with true quality q_i. We also assume that $B(\bar{a})$, from Eq. (2), corresponds to reproductive value. These reproductive values represent payoffs in the standard sense of evolutionary game theory. The meaning of the perceived quality p_i is that the individual perceives the cost $K(a_i, p_i)$, in the sense of rewards influencing learning.

Let $a_i^*(p)$ be a Nash equilibrium where the true qualities in Eq. (13) are replaced by perceived qualities, thus satisfying

$$\frac{1}{g}B'(\bar{a}^*(p)) = \frac{\partial K}{\partial a_i}(a_i^*(p), p_i), \tag{15}$$

for $i = 1, \ldots, g$. If the true qualities of group members are q_i, an evolutionary equilibrium for the perceived qualities p_i should satisfy (see SI)

$$\begin{aligned}\frac{dW_i}{dp_i} &= \left[\frac{\partial K}{\partial a_i}(a_i^*(p), p_i) - \frac{\partial K}{\partial a_i}(a_i^*(p), q_i)\right]\frac{\partial a_i^*(p)}{\partial p_i} \\ &\quad + \frac{1}{g}B'(\bar{a}^*(p))\sum_{j \neq i}\frac{\partial a_j^*(p)}{\partial p_i} = 0.\end{aligned} \tag{16}$$

From this it follows that $p_i = q_i$ is not an evolutionary equilibrium for $g > 1$, because the expression in the square bracket is then zero and the other term is negative, because $\partial a_j^*/\partial p_i < 0$ for $j \neq i$. This shows that an individual could gain fitness by lowering its perceived quality from $p_i = q_i$ to $p_i = q_i + d_i$ with $d_i < 0$.

In such a case, an individual with true quality q_i will perceive the cost $K(a_i, p_i) = K(a_i, q_i + d_i)$. For our special case of Eq. (3), this means that the individual perceives an extra cost, or penalty, $K_{12} d_i a_i$ of the investment a_i. The solution to Eq. (16) for the special case can be written as

$$p_i^* - q_i = \beta_0 + \beta_1 \bar{q}, \tag{17}$$

which is worked out in the SI, with β_0 and β_1 given in Equation (S30). For $g = 1$ one sees from Equation (S30) that $\beta_0 = \beta_1 = 0$, so that $p_1^* = q_1$ is the solution.

Individual-based simulations. For individual-based simulation of the actor-critic learning dynamics, we constructed populations of individuals, each with a randomly assigned quality, split into groups of size g. For ease of interpretation, qualities were drawn from a small set of values for q, for instance $q_i \in \{0, 1\}$ in Fig. 2. In this population, the learning dynamics follows Eqs (5–10) over rounds $t = 1, \ldots, T$. The aim of the simulations is to compare the outcome of learning with the one-shot Nash equilibrium predictions from Eq. (14). For evolutionary simulations, over many generations, we implemented discrete, non-overlapping generations and assumed individuals to be hermaphrodites with one diploid locus additively determining the trait $d_i = p_i - q_i$. The time sequence of events for evolutionary simulations was as follows: (*i*) random sorting of newborn individuals into groups and assignment of random true qualities; (*ii*) learning dynamics over T rounds, with the perceived quality of an individual given as

$$p_i = q_i + d_i, \tag{18}$$

where d_i is the individual's genetically determined trait; (*iii*) assignment of a Darwinian payoff to each individual, computed as the individual's average payoff over the rounds, based on its true quality; and (*iv*) formation of the next generation through mating, including mutation, with the probability of being chosen as parent being proportional to an individual's payoff.

Data availability

Source code for the individual-based simulations is available at GitHub, together with instruction for compilation on a Linux operating system, and with example input files: https://github.com/oleimar/pggsim. The R code and individual-based simulation output used to generate the figures are available from the corresponding author on reasonable request.

Received: 30 July 2019; Accepted: 22 October 2019;
Published online: 08 November 2019

References

1. Staddon, J. *Adaptive Behavior and Learning* (Cambridge University Press, Cambridge, 2016).
2. Sutton, R. S. & Barto, A. G. *Reinforcement Learning: An Introduction*, *Second Edition* (MIT Press, Cambridge, MA, 2018).
3. McNamara, J. & Leimar, O. Variation and the response to variation as a basis for successful cooperation. *Philos. Transactions Royal Soc. B: Biol. Sci.* **365**, 2627–2633 (2010).
4. McNamara, J. M. Towards a richer evolutionary game theory. *J. Royal Soc. Interface* **10** (2013).
5. Gigerenzer, G. & Selten, R. *Bounded Rationality: The Adaptive Toolbox* (MIT Press, Cambridge, MA, 1999).
6. Erev, I. & Roth, A. E. Maximization, learning, and economic behavior. *Proc. Natl. Acad. Sci.* **111**, 10818–10825 (2014).
7. McNamara, J. M., Gasson, C. E. & Houston, A. I. Incorporating rules for responding into evolutionary games. *Nat.* **401**, 368–371 (1999).
8. Bush, R. R. & Mosteller, F. *Stochastic Models for Learning* (John Wiley & Sons Inc., New York, 1955).
9. Lütkepohl, H. *New Introduction to Multiple Time Series Analysis* (Springer, Berlin, 2005).
10. Pavlov, I. P. *Conditioned Reflexes: An Investigation of the Physiological Activity of the Cerebral Cortex* (Oxford University Press, Oxford, 1927).
11. Enquist, M., Lind, J. & Ghirlanda, S. The power of associative learning and the ontogeny of optimal behaviour. *Royal Soc. Open Sci.* **3**, 160734 (2016).
12. Kuhn, H. W. *et al*. The work of John Nash in game theory - Nobel Seminar, December 8, 1994. *J. Econ. Theory* **69**, 153–185 (1996).
13. Weibull, J. W. *Evolutionary Game Theory* (MIT Press, Cambridge, MA, 1995).
14. Fudenberg, D. & Levine, D. K. *The Theory of Learning in Games* (MIT Press, Cambridge, MA, 1998).
15. Sandholm, W. H. *Population Games and Evolutionary Dynamics* (MIT Press, Cambridge, MA, 2010).
16. Harley, C. B. Learning the evolutionarily stable strategy. *J. Theor. Biol.* **89**, 611–633 (1981).
17. Maynard Smith, J. *Evolution and the Theory of Games* (Cambridge University Press, Cambridge, 1982).
18. Dridi, S. & Lehmann, L. On learning dynamics underlying the evolution of learning rules. *Theor. Popul. Biol.* **91**, 20–36 (2014).
19. Rescorla, R. A. & Wagner, A. R. A theory of Pavlovian conditioning: variations in the effectiveness of reinforcement and nonreinforcement. In Black, A. H. & Prokasy, W. F. (eds) *Classical Conditioning II: Current Research and Theory*, 64–99 (Appleton-Century-Crofts, New York, 1972).
20. Montague, P. R., Dayan, P. & Sejnowski, T. J. A framework for mesencephalic dopamine systems based on predictive Hebbian learning. *J. Neurosci.* **16**, 1936–47 (1996).
21. Gerstner, W., Lehmann, M., Liakoni, V., Corneil, D. & Brea, J. Eligibility traces and plasticity on behavioral time scales: experimental support of neoHebbian three-factor learning rules. *Front. Neural Circuits* **12**, 1–16 (2018).
22. Price, G. R. Selection and covariance. *Nat.* **227**, 520–521 (1970).
23. Roth, A. E. & Erev, I. Learning in extensive-form games: experimental data and simple dynamic models in the intermediate term. *Games Econ. Behav.* **8**, 164–212 (1995).
24. Erev, I. & Roth, A. E. Predicting how people play games: reinforcement learning in experimental games with unique, mixed strategy equilibria. *Am. Econ. Rev.* **88**, 848–881 (1998).
25. Tversky, A. & Kahneman, D. Evidential impact of base rates. In Kahneman, D., Slovic, P. & Tversky, A. (eds) *Judgment under Uncertainty: Heuristics and Biases*, chap. 10, 153–160 (Cambridge University Press, Cambridge, 1982).
26. Mendl, M., Burman, O. H. P. & Paul, E. S. An integrative and functional framework for the study of animal emotion and mood. *Proc. Royal Soc. B* **277**, 2895–2904 (2010).
27. Marshall, J. A. R., Trimmer, P. C., Houston, A. I. & McNamara, J. M. On evolutionary explanations of cognitive biases. *Trends Ecol. Evol.* **28**, 469–473 (2013).
28. Güth, W. An evolutionary approach to explaining cooperative behavior by reciprocal incentives. *Int. J. Game Theory* **24**, 323–344 (1995).
29. Robson, A. J. & Samuelson, L. The evolutionary foundations of preferences. In Benhabib, J., Bisin, A. & Jackson, M. (eds) *Handbook of Social Economics*, vol. 1B, chap. 7, 221–310 (Elsevier B.V., Amsterdam, 2011).
30. Houston, A. I. & McNamara, J. M. *Models of Adaptive Behaviour* (Cambridge University Press, Cambridge, 1999).
31. Caswell, H. *Matrix Population Models*, *Second Edition* (Sinauer Associates, Inc., Sunderland, MA, 2001).
32. Heifetz, A., Shannon, C. & Spiegel, Y. What to maximize if you must. *J. Econ. Theory* **133**, 31–57 (2007).
33. Alger, I. & Weibull, J. W. Evolutionary models of preference formation. *Annu. Rev. Econ.* **11**, 329–354 (2019).
34. Sih, A., Bell, A. M., Johnson, J. C. & Ziemba, R. E. Behavioral syndromes: an integrative overview. *The Q. Rev. Biol.* **79**, 241–277 (2004).
35. Wolf, M., van Doorn, G. S., Leimar, O. & Weissing, F. J. Life-history trade-offs favour the evolution of animal personalities. *Nat.* **447**, 581–584 (2007).
36. Rosen, J. B. Existence and uniqueness of equilibrium points for concave N-person games. *Econom.* **33**, 520–534 (1965).
37. Miller, K. S. On the inverse of the sum of matrices. *Math. Mag.* **54**, 67–72 (1981).

Acknowledgements

We thank Tim Fawcett, Andy Higginson and Alasdair Houston for helpful comments. This work was supported by a grant (2018-03772) from the Swedish Research Council to O.L.

Author contributions

Wrote the paper: O.L. with input from J.M.M. Conceived and performed the modelling: O.L. with input from J.M.M. Designed the software for the model analysis: O.L.

Competing interests
The authors declare no competing interests.

Additional information
Supplementary information is available for this paper at https://doi.org/10.1038/s41598-019-52781-7.

Correspondence and requests for materials should be addressed to O.L.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.